

# Mitochondrial DNA variation and language replacements in the Caucasus

Ivane Nasidze\*† and Mark Stoneking†

*Department of Anthropology, Pennsylvania State University, University Park, PA 16802, USA*

Sequences of the first hypervariable segment of the mitochondrial DNA (mtDNA) control region were obtained from 353 individuals representing nine groups and four major linguistic families (Indo-European, Altaic and North and South Caucasian) of the Caucasus region. The diversity within and between Caucasus populations exceeded the diversity within Europe, but was less than that in the Near East. Caucasus populations occupy an intermediate position between European and Near Eastern populations in tree and principal coordinate analyses, suggesting that they are either ancestral to European populations or derived via admixture from European and Near Eastern populations. The genetic relationships among Caucasus populations reflect geographical rather than linguistic relationships. In particular, the Indo-European-speaking Armenians and Altaic-speaking Azerbaijanians are most closely related to their nearest geographical neighbours in the Caucasus, not their linguistic neighbours (i.e. other Indo-European or Altaic populations). The mtDNA evidence thus suggests that the Armenian and Azerbaijani languages represent instances of language replacement that had little impact on the mtDNA gene pool.

**Keywords:** mitochondrial DNA; human variation; Caucasus; language replacement

## 1. INTRODUCTION

The Caucasus (the area between the Caspian and Black Seas) exhibits tremendous linguistic diversity, with four major language families represented (North Caucasian, South Caucasian, Indo-European and Altaic). A significant geographical barrier, the Caucasus Mountains, which extend for 1200 km, divides the region into the North and South Caucasus sub-regions. This region thus offers the opportunity for examining the impact of linguistic diversity and geographical barriers on genetic diversity in addition to potential insights that might be gained into European–Asian migrations.

Little is known about the prehistory of the Caucasus. While the oldest, securely dated Neolithic sites are *ca.* 6000–7000 years old (Muskhelishvili 1977), many sites have not been dated or thoroughly studied. Linguistic evidence concerning Caucasus population relationships is similarly equivocal. There is debate as to whether or not the North and South Caucasian languages together form a genetic unit (Ruhlen 1991). For example, Renfrew (1992) considered both families to be relics of the initial dispersal of humans from Africa more than 15 000 years ago, while the Armenian (Indo-European) and Azerbaijani (Altaic) languages spread into the Caucasus more recently by the process of elite dominance. Nichols (1997) argued instead that the South Caucasian languages entered the Caucasus recently and that the Armenian language is a remnant of a formerly more widespread language.

Previous studies of classical genetic markers in the Caucasus (Barbujani *et al.* 1994*a,b*) have found a correspondence between genetic and linguistic relationships at

a local level only within the same language. When Caucasus populations that spoke different languages were analysed, their genetic relationships correlated more strongly with geographical relationships than with linguistic relationships. Overall, the analyses of classical genetic markers were found to be in agreement with a single ancient origin of Caucasus populations, followed by subdivision along geographical and linguistic lines.

We have previously analysed eight *Alu* insertion polymorphisms in six populations from the Caucasus (Nasidze *et al.* 2001). The Caucasus populations exhibit high levels of between-population differentiation that are almost as large as for worldwide populations. Neither geographical nor linguistic relationships appear to explain their genetic relationships. Instead, it appears as if they have been small and relatively isolated and, hence, that genetic drift has been the dominant influence on the genetic structure of Caucasus populations.

Mitochondrial DNA (mtDNA) is a useful and informative genetic marker for investigating human population history. However, to date only limited information is available concerning mtDNA diversity in the Caucasus (Macaulay *et al.* 1999; Comas *et al.* 2000). We therefore undertook a comprehensive analysis of mtDNA diversity in nine populations. Among the questions we addressed in this study are the following.

- (i) Are the Caucasus populations genetically heterogeneous?
- (ii) Does genetic differentiation correlate with linguistic differentiation?
- (iii) Do major geographical boundaries such as the Caucasus Mountains demonstrably influence the patterns of genetic variation and differentiation in the Caucasus?
- (iv) How are the Caucasus populations genetically related to European and Near Eastern populations?

\*Author for correspondence (nasidze@eva.mpg.de).

†Present Address: Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany.

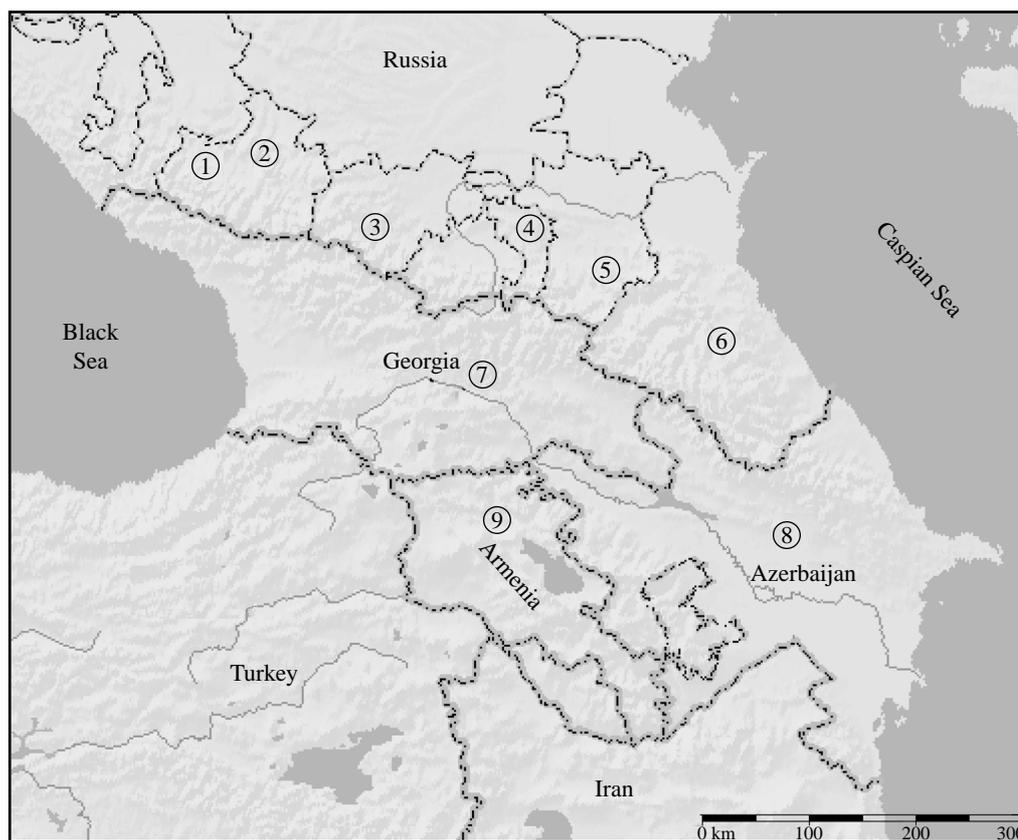


Figure 1. Map of the Caucasus region indicating the locations of the populations used in this study. 1, Cherkessians; 2, Abazinians; 3, Kabardinians; 4, Ingushians; 5, Chechenians; 6, Darginians; 7, Georgians; 8, Azerbaijanians; 9, Armenians.

## 2. MATERIAL AND METHODS

### (a) *Samples*

A total of 353 peripheral blood samples from unrelated individuals were collected from the following nine autochthonous groups (figure 1), which represent all major linguistic families in the Caucasus: Georgians (South Caucasian), Armenians (Indo-European), Azerbaijanians (Altaic) and Abazinians, Cherkessians, Kabardinians, Ingushians, Chechenians and Darginians (North Caucasian). Informed consent and information about birthplace, parents and grandparents was obtained from all donors. Published mtDNA HV1 sequences were also used from 50 Adygheians (Macaulay *et al.* 1999) from the north-west Caucasus. Another sample of 45 Georgian HV1 sequences that recently became available (Comas *et al.* 2000) does not differ in any respect from our Georgian sample (data not shown) and, hence, was not included in order to avoid weighting the results too heavily on one population.

### (b) *DNA extraction and sequencing*

Total genomic DNA was extracted from whole blood using an IsoQuick DNA extraction kit (Orca Research, Inc., Bothell, WA, USA). Primers L15996 and H16401 (Vigilant *et al.* 1989) were used for amplifying the first hypervariable segment (HV1) of the mtDNA control region, as described previously (Redd *et al.* 1995). These primers were used for determining sequences for both strands of a polymerase chain reaction's products with a DNA Sequencing Kit (Applied Biosystems, Foster City, CA, USA), following the protocol recommended by the supplier and an ABI 373 automated DNA sequencer. The sequences have

been submitted to the HvrBase database (Burckhardt *et al.* 1999).

### (c) *Statistical analysis*

The basic parameters of molecular diversity and population genetic structure (including analyses of molecular variance, AMOVA) were calculated using the computer program Arlequin 2.0 (Schneider *et al.* 2000). The statistical significance of  $F_{ST}$ - and  $\Phi_{ST}$ -values was estimated by permutation analysis using 10 000 permutations. The statistical significance of the correlation between geographical distance and genetic distance matrices was evaluated by the Mantel test with 1000 permutations using the Permute! 3.4 program (Legendre *et al.* 1994). Mismatch distributions of the number of nucleotide substitutions within populations were analysed as described previously (Rogers & Harpending 1992; Harpending *et al.* 1993). Genetic distances between individual sequences were calculated as described elsewhere (Calafell *et al.* 1996) and neighbour-joining trees were produced with programs in PHYLIP3.5c (Felsenstein 1993). The STATISTICA package (StatSoft, Inc., Tulsa, OK, USA) was used for principal coordinate (PC) analysis, with the ordination derived from the covariance matrix of raw haplotype frequencies.

Mitochondrial DNA HV1 sequences from 106 Basques (Bertranpetit *et al.* 1995; C  rte-Real *et al.* 1996), 101 Britons (Piercy *et al.* 1993), 69 Sardinians and 42 Middle Easterners (DiRienzo & Wilson 1991), 72 Spaniards (C  rte-Real *et al.* 1996; Richards *et al.* 1996), 96 Turks (Calafell *et al.* 1996; Comas *et al.* 1996; Richards *et al.* 1996), 45 Israeli Drusi (Macaulay *et al.* 1999), 29 Kurds (Comas *et al.* 2000) and 98 Indians (Mountain

Table 1. Parameters summarizing some characteristics of the mtDNA HV1 sequence diversity among the Caucasus, European and Near Eastern groups

population	sample size	number of haplotypes	haplotype diversity	nucleotide diversity	% reference sequence
Abazians	23	19	0.980 ± 0.020	0.014 ± 0.008	13.0
Armenians	42	35	0.980 ± 0.014	0.014 ± 0.008	14.3
Azerbaijanians	41	37	0.995 ± 0.007	0.014 ± 0.008	4.9
Chechenians	23	18	0.972 ± 0.022	0.012 ± 0.007	13.0
Cherkessians	44	37	0.986 ± 0.010	0.015 ± 0.008	11.4
Darginians	37	27	0.974 ± 0.015	0.014 ± 0.008	13.5
Georgians	57	40	0.971 ± 0.014	0.014 ± 0.008	15.8
Ingushians	35	26	0.970 ± 0.018	0.013 ± 0.007	14.3
Kabardinians	51	36	0.975 ± 0.011	0.013 ± 0.007	11.8
Adygheian	50	32	0.953 ± 0.018	0.014 ± 0.007	16.0
<b>Caucasus (total)</b>	<b>403</b>	<b>238</b>	<b>0.980 ± 0.004</b>	<b>0.014 ± 0.008</b>	<b>12.9</b>
Basques	106	53	0.936 ± 0.018	0.010 ± 0.006	23.6
British	101	68	0.973 ± 0.010	0.012 ± 0.007	14.7
Sardinians	69	45	0.944 ± 0.022	0.012 ± 0.007	24.6
Spanish	72	42	0.939 ± 0.022	0.011 ± 0.006	14.1
Kurds	29	22	0.985 ± 0.028	0.012 ± 0.007	20.7
Turkish	96	79	0.988 ± 0.006	0.015 ± 0.008	8.3
Drusi	45	26	0.948 ± 0.019	0.013 ± 0.007	11.1
Middle East	42	41	0.999 ± 0.006	0.019 ± 0.010	0.0
Indians	98	44	0.961 ± 0.010	0.018 ± 0.009	0.9

*et al.* 1995) were used for some analyses for comparison with the Caucasus sequences. Other European and central Asian populations were also included in the preliminary analyses, but did not alter any conclusions based on the above populations only.

### 3. RESULTS

#### (a) Sequence variability

A total of 377 bp of the mtDNA HV1 region, comprising nucleotide positions 16 024–16 400 (Anderson *et al.* 1981), were determined for 353 individuals from the nine Caucasus groups. The proportion of transitions varied from 84% of the polymorphic sites found among Azerbaijanians to 100% among Darginians. At one position both transition and transversion mutations occurred in the Azerbaijanians (position 16 304) and Darginians (position 16 368), while two additional sites (positions 16 153 and 16 247) presented both types of mutations in Georgians.

Subsequent analyses were restricted to 365 bp (nucleotide positions 16 024–16 388) of HV1 for the purpose of comparing the sequences reported here with published data. The nucleotide diversity ranged from 0.010 to 0.015 in the various Caucasus groups (table 1), with an overall average of 0.014, while the haplotype diversity ranged from 0.953 to 0.995. On average, the nucleotide and haplotype diversities were slightly higher in the Caucasus populations than in the European populations, but were lower than in the Turkish, Middle East and Indian populations. A non-parametric (Mann–Whitney) test indicated that both the haplotype and nucleotide diversities were significantly higher ( $p < 0.05$ ) among the Caucasus populations than among the European populations.

There were 238 distinct haplotypes among the 403 Caucasus individuals, with the only haplotype shared by

all Caucasus populations being the published reference sequence (Anderson *et al.* 1981). The frequency of this sequence varied between 11 and 16% in the Caucasus populations, except for the Azerbaijanian population, in which it was only 4.9% (table 1). European populations tended to have somewhat higher frequencies of this sequence (14–25%), whereas the frequency was lower in the Turkish and Indian populations (8 and 1%, respectively). Thus, the Caucasus populations appeared to be similar to European populations, but with somewhat higher mtDNA diversity.

#### (b) Phylogenetic analysis

The phylogenetic tree of the 238 unique sequences from the Caucasus had an approximately star-like pattern and revealed no clear clustering of the various groups, with sequences from all groups scattered throughout the tree (figure 2). The robustness of the tree was estimated by performing 1000 bootstrap replications. The lengths of the most peripheral branches were significantly different from zero by this analysis. However, the central branches were not significantly different from zero (data not shown). This is a clear indication of the star-like shape of the tree, which is consistent with an expanding population, as investigated further in § 3(c).

#### (c) Pairwise nucleotide difference distributions

The mean number of pairwise nucleotide differences was fairly uniform across the different Caucasus groups, ranging from 4.40 to 5.35 (table 2). These estimates were towards the upper limit of the range of mean pairwise differences found in European populations (3.15–5.03) (Comas *et al.* 1997), but lower than those for Turkish (5.45) and Middle Eastern (7.08) groups. The mismatch distributions for the Caucasus groups were all approximately bell shaped (figure 3), suggesting prehistoric

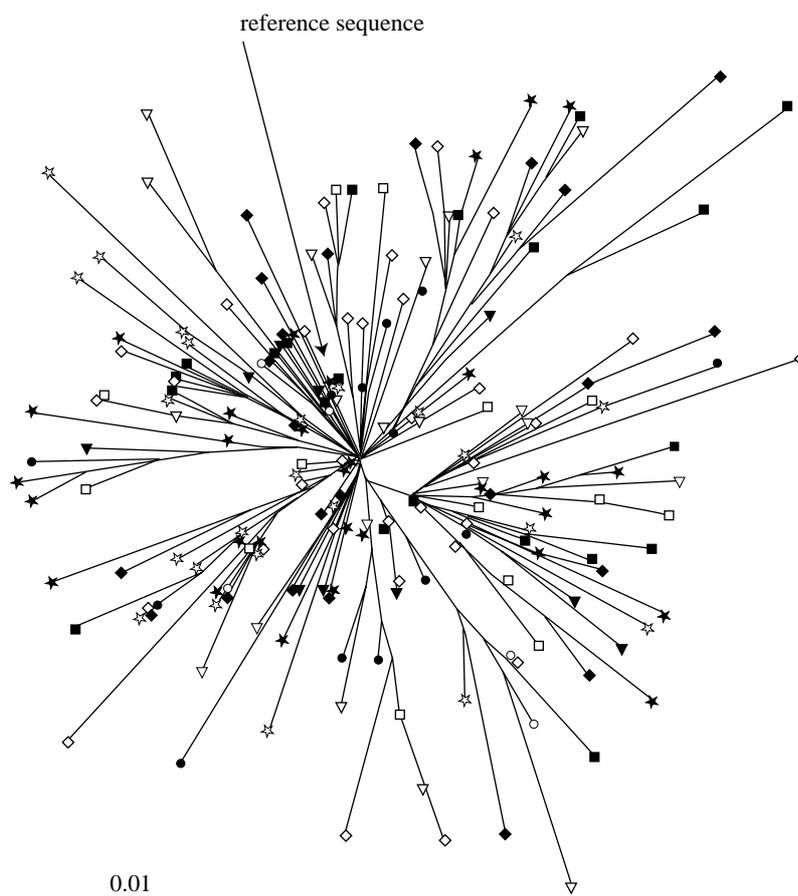


Figure 2. Neighbour-joining tree for the 238 distinct mtDNA sequences in the Caucasus groups constructed from genetic distances based on the Kimura two-parameter model with a 10:1 ratio of transitions to transversions. Inverted filled triangles, Adygheians; inverted open triangles, Abazinians; filled squares, Armenians; filled diamonds, Azerbaijanians; open squares, Chechenians; open diamonds, Cherkessians; filled circles, Darginians; open circles, Georgians; filled stars, Ingushians; open stars, Kabardinians.

Table 2. Summary statistics for the pairwise nucleotide difference distributions for the Caucasus groups

(The expansion times were calculated assuming an evolutionary rate of 33% per million years (Ward *et al.* 1991) for 360 bp of HV1. \* $p < 0.05$ .)

population	mean number of mismatches	variance number of mismatches	raggedness index	Tajima's $D$	$\tau$	expansion time (Kyr)
Abazinians	5.19	7.84	0.010	-2.02*	3.44	29
Adygheians	4.98	5.48	0.017	-1.55	5.56	47
Armenians	5.22	6.99	0.011	-2.18*	4.95	42
Azerbaijanians	5.17	4.47	0.015	-2.13*	5.26	44
Chechenians	4.40	3.85	0.031	-1.67*	4.52	38
Cherkessians	5.35	6.88	0.009	-1.98*	4.60	39
Darginians	5.10	8.62	0.008	-2.04*	3.70	31
Georgians	5.16	8.17	0.008	-1.99*	3.91	33
Ingushians	4.75	4.99	0.015	-1.57	5.22	44

population expansions. The raggedness statistic for the Caucasus mismatch distributions varied from 0.008 to 0.031 (table 2); values of  $r$  less than 0.05 are also indicative of prehistoric population expansions (Harpending *et al.* 1993). This demographic scenario was reinforced by Tajima's (1989)  $D$ -statistic, which was negative in all of the Caucasus groups and significantly so in all but the Adygheians and Ingushians (table 2); negative values of

$D$ , together with bell-shaped mismatch distributions, are signatures of population expansions (Aris-Brosou & Excoffier 1996).

Assuming that the mismatch distributions did therefore reflect past population expansions, the method of Schneider & Excoffier (1999) was used for estimating  $\tau$ , the time of population expansion in units of mutational time (table 2). The estimated  $\tau$ -values varied from 3.44 to

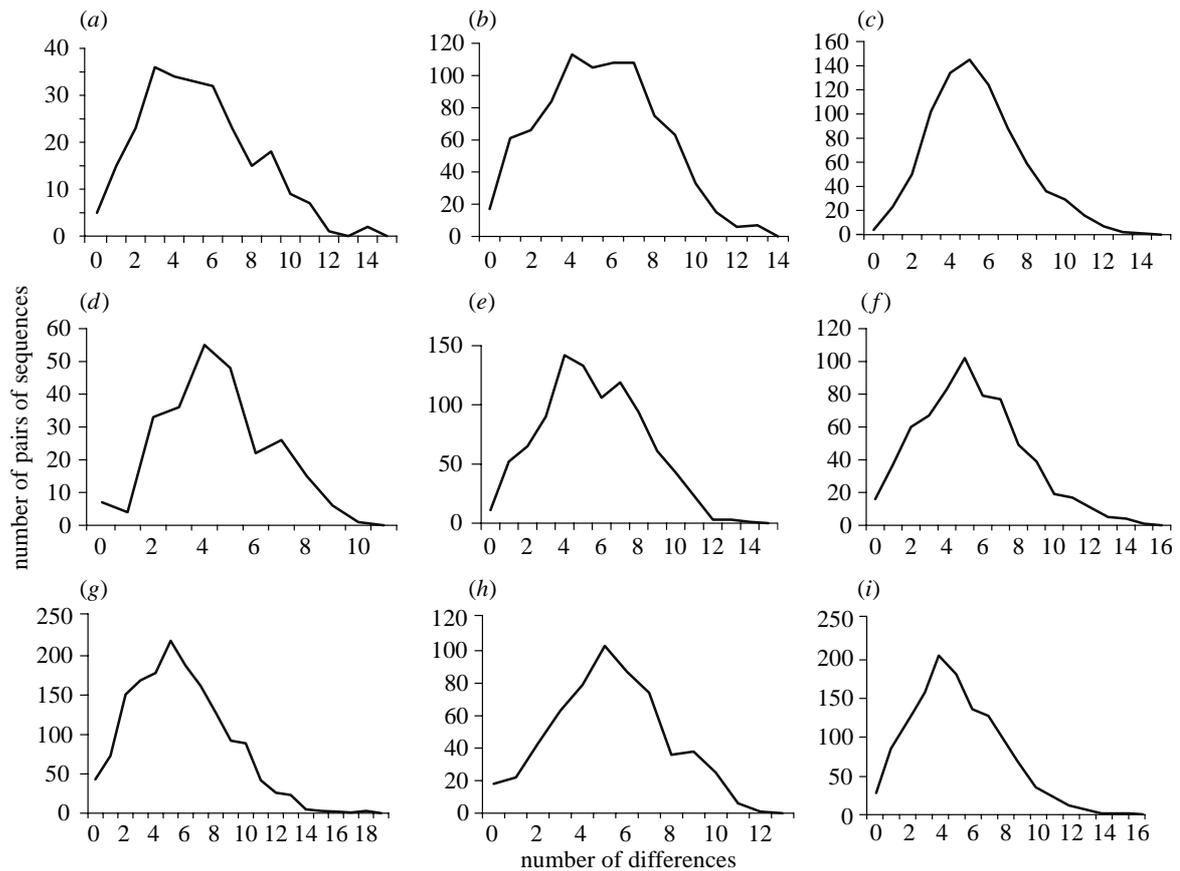


Figure 3. Pairwise nucleotide difference distributions for the Caucasus groups. (a) Abazinians, (b) Armenians, (c) Azerbaijanians, (d) Chechenians, (e) Cherkessians, (f) Darginians, (g) Georgians, (h) Ingushians and (i) Kabardinians.

5.56, which correspond to estimated expansion times of 29 000–47 000 years ago, assuming a rate of human mtDNA divergence of 33% per million years (Ward *et al.* 1991).

#### (d) *Caucasus population structure*

AMOVA showed that, when the ten populations were treated as a single group, 98.6% of the total variance was within populations and 1.4% (which was statistically significant at  $p < 0.001$ ) was between populations (table 3). The  $F_{ST}$ - and  $D_A$ -values (table 4) were highest between Ingushians and the other Caucasus groups (0.025 and 0.116, respectively, compared to average values of 0.009 and 0.044, respectively, among the other nine groups).  $F_{ST}$ -values with associated probabilities less than 0.01 were considered nominally significant and are highlighted in table 4; four  $F_{ST}$ -values involving Ingushians had a significance level of  $p < 0.01$ , while the remaining four  $F_{ST}$ -values with a significance level of  $p < 0.01$  all involved Adygheians (table 4). The average  $F_{ST}$ - and  $D_A$ -values were smaller within the Southern Caucasus ( $F_{ST} = 0.007$  and  $D_A = 0.034$ ) and the Northern Caucasus ( $F_{ST} = 0.010$  and  $D_A = 0.048$ ) than those between the Southern and Northern Caucasus ( $F_{ST} = 0.015$  and  $D_A = 0.071$ ), providing some evidence for genetic structuring corresponding to the geographical sub-regions.

#### (e) *Relationships with other populations*

The pairwise  $F_{ST}$ - and  $D_A$ -values between the Caucasus and European, Near Eastern and Indian groups indicated

Table 3. *AMOVA results according to different classifications*

(Classifications: linguistic 1 (Turkic, Indo-European and South (Kartvelian) and North Caucasian), linguistic 2 (Turkic, Indo-European and Caucasian), linguistic 3 (Turkic, Indo-European and South Caucasian (Kartvelian) and North Caucasian subgroups (Abkhazo-Adygheian and Central and Dagestanian subgroups) and geographical (South and North Caucasus). All between-group and between-population–within-group values were significantly different from zero based on 10 000 permutations.)

classification	percentage of variation		
	between groups	between populations–within groups	within populations
individual groups	0.27	—	99.73
linguistic 1	0.15	1.15	98.70
linguistic 2	0.33	1.10	98.56
linguistic 3	0.41	0.88	98.71
geographic	0.42	1.02	98.56

that the Caucasus and Europe are most closely associated (table 4). The average  $F_{ST}$ - and  $D_A$ -values between the Caucasus and the Near East were approximately three times larger ( $F_{ST} = 0.060$  and  $D_A = 0.195$ ) than those between the Caucasus and Europe ( $F_{ST} = 0.019$  and  $D_A = 0.068$ ). Moreover, the population tree and PC plot (figure 4) clearly grouped the Caucasus and European

Table 4. *Pairwise  $F_{ST}$ -values between the Near East, European and Caucasus groups (below diagonal) and  $D_A$ -values (above diagonal)*(  $F_{ST}$ -values marked with asterisks indicate a significance level of  $p < 0.01$ .)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Cherkessians	—	-0.006	-0.009	0.148	0.078	0.020	0.023	0.071	0.058	0.112	0.104	0.103	0.065	0.088	0.077	0.076	0.281	0.314	0.486
Abazinians	-0.001	—	0.043	0.048	-0.008	-0.015	0.006	0.073	0.030	-0.024	0.026	0.038	0.003	0.019	0.036	0.036	0.197	0.259	0.502
Kabardinians	-0.002	-0.009	—	0.055	0.071	0.022	0.007	0.060	0.038	0.017	0.033	0.045	0.019	0.018	-0.007	-0.007	0.241	0.241	0.458
Ingushians	0.031*	0.011	0.012	—	0.152	0.091	0.136	0.204	0.132	0.079	0.136	0.175	0.123	0.160	0.131	0.131	0.364	0.396	0.535
Chechenians	0.016	0.002	0.015	0.033	—	0.048	0.021	0.067	0.031	0.110	0.090	0.081	0.014	0.077	0.144	0.176	0.168	0.281	0.513
Darginians	0.004	0.003	0.005	0.020	0.010	—	0.041	0.083	0.056	0.062	0.076	0.045	0.043	0.073	0.092	0.092	0.231	0.223	0.439
Georgians	0.005	0.001	0.001	0.028*	0.003	0.009	—	0.052	0.032	0.090	0.061	0.070	0.039	0.048	0.048	0.047	0.218	0.269	0.433
Azerbaijanians	0.015	0.015	0.013	0.042*	0.013	0.017	0.011	—	0.019	0.143	0.080	0.111	0.077	0.071	0.088	0.087	0.291	0.300	0.585
Armenians	0.012	0.006	0.008	0.028*	0.006	0.011	0.007	0.004	—	0.131	0.087	0.070	0.037	0.071	0.038	0.037	0.1764	0.250	0.486
Adygheians	0.023*	-0.005	0.004	0.016	0.021	0.013	0.018*	0.029*	0.026*	—	0.072	0.039	0.054	0.072	0.070	0.070	0.196	0.223	0.654
Basques	0.032*	0.017	0.012	0.043*	0.034*	0.026*	0.019*	0.026*	0.028*	0.022*	—	0.045	0.028	-0.021	0.059	0.059	0.330	0.295	0.630
Sardinians	0.024*	0.011	0.011	0.040*	0.020	0.011	0.016*	0.025*	0.016*	0.009	0.014*	—	0.008	0.043	0.044	0.043	0.145	0.127	0.612
Britains	0.015*	0.003	0.005	0.028*	0.003	0.011	0.009	0.018*	0.009	0.013	0.008	0.002	—	0.019	0.042	0.042	0.145	0.182	0.553
Spanish	0.025*	0.012	0.006	0.046*	0.027	0.022*	0.013	0.021*	0.021*	0.020*	-0.007	0.012	0.004	—	0.042	0.099	0.294	0.275	0.586
Kurds	0.016*	0.008	-0.002	0.029*	0.032*	0.020*	0.009	0.018*	0.008	0.014	0.023*	0.011	0.010	0.016*	—	0.065	0.214	0.158	0.485
Turkish	0.094*	0.074*	0.079*	0.099*	0.087*	0.077*	0.087*	0.099*	0.092*	0.080*	0.138*	0.083*	0.098*	0.021*	0.010*	—	0.436	0.427	0.689
Middle Easterns	0.047*	0.027	0.042*	0.058*	0.023	0.037*	0.038*	0.047*	0.029*	0.033*	0.081*	0.029*	0.031*	0.065*	0.033*	0.012*	—	0.050	0.620
Drusi	0.064*	0.053*	0.050*	0.080*	0.058*	0.046*	0.054*	0.060*	0.051*	0.045*	0.081*	0.029*	0.040*	0.071*	0.034*	0.018	0.009	—	0.588
Indians	0.078*	0.077*	0.076*	0.085*	0.078*	0.070*	0.071*	0.092*	0.078*	0.103*	0.124*	0.104*	0.097*	0.108*	0.076*	0.064*	0.090*	0.094*	—

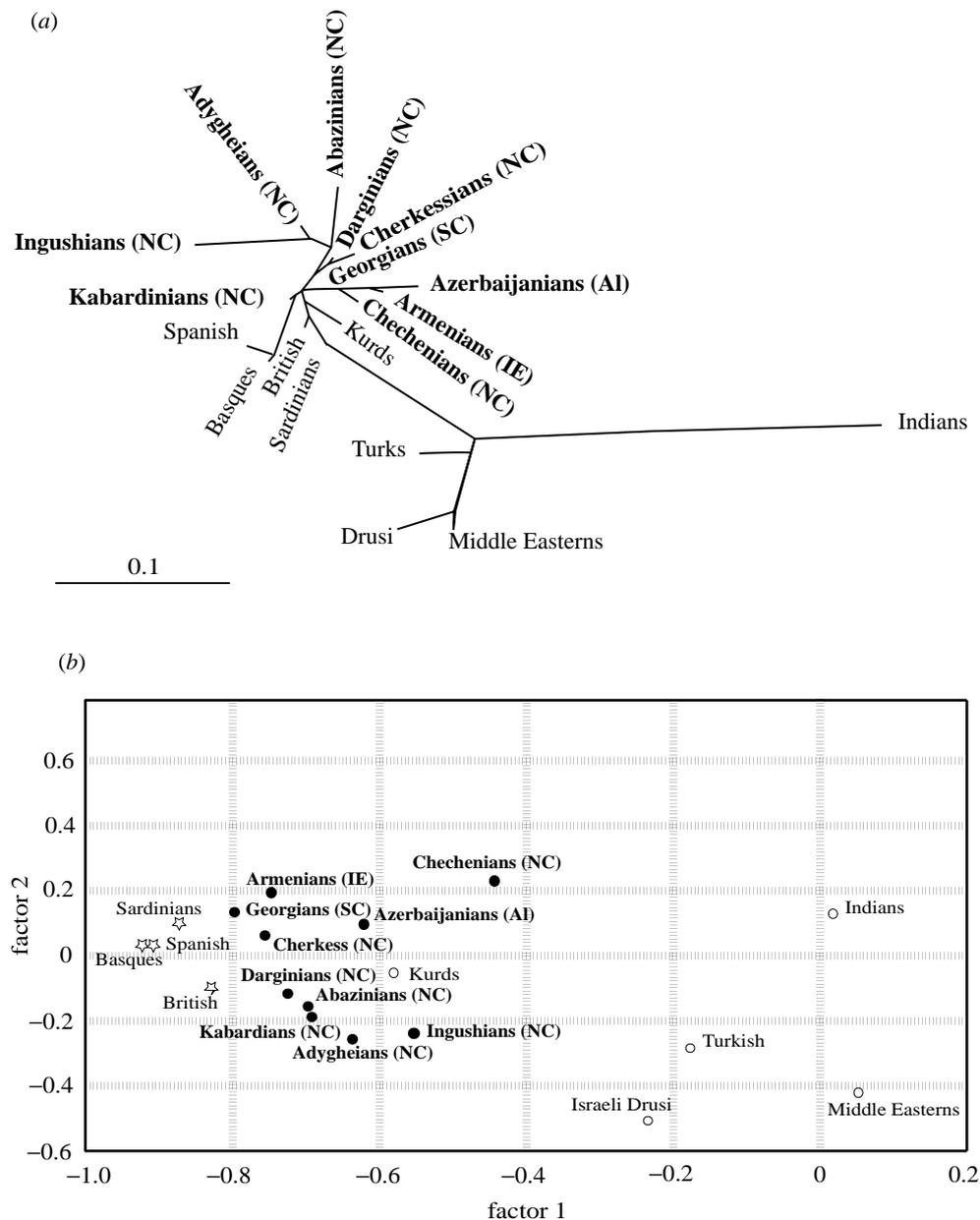


Figure 4. (a) Population tree and (b) PC plot illustrating the relationships between the Caucasus, European, Near Eastern and Indian populations. The suffixes for the Caucasus groups (given in bold) indicate language family: NC, North Caucasian; SC, South Caucasian (Kartvelian); IE, Indo-European; AI, Altaic. The neighbour-joining tree of the populations was constructed from the  $D_A$ -values, which were calculated as described in § 2.

groups together, apart from the Near Eastern and Indian groups. Figure 4 also shows the greater diversity of Caucasus groups relative to European groups, which is also evident in the  $F_{ST}$ - and  $D_A$ -values (table 4); the average  $F_{ST}$ - and  $D_A$ -values were bigger among the Caucasus groups (0.012 and 0.058, respectively) than among the European groups (0.006 and 0.020, respectively).

(f) **Genetic, linguistic and geographical relationships of Caucasus groups**

The geographical and linguistic structures of Caucasus mtDNAs were investigated by the AMOVA procedure. When the Caucasus groups were clustered by various geographical or linguistic criteria, the proportion of the genetic variance that was due to differences between

groups was actually less than that due to differences between subpopulations within groups, no matter how the groups were apportioned (table 3). Nonetheless, even though 98–99% of the total genetic variance was within subpopulations, the between-group proportion was significantly different from zero for all comparisons. However, the AMOVA results (table 3) indicated that the presumed linguistic barriers in the Caucasus are not reflected in the patterns of mtDNA diversity.

The  $F_{ST}$ - and  $D_A$ -values (table 4) also gave some indication that the patterns of mtDNA diversity do not reflect the linguistic relationships of the Caucasus groups. The Indo-European-speaking Armenians had an average  $F_{ST}$ -value of 0.014 when compared with other Indo-European-speaking groups, which is nearly identical to

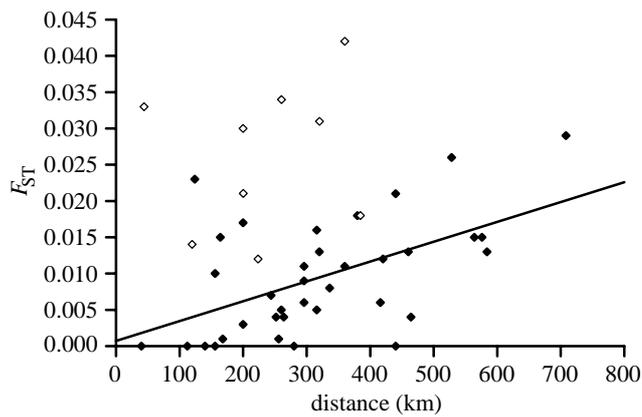


Figure 5. Relationships between geographical distances (measured as a straight line) and genetic ( $F_{ST}$ ) distances between pairs of Caucasus groups. Open symbols indicate population pairs that include Ingushians, illustrating that they are outliers with higher  $F_{ST}$ -values and, hence, were not included in the regression analysis.

that between Armenians and Caucasian-speaking groups (average  $F_{ST} = 0.013$ ). The Altaic-speaking Azerbaijanians were much more similar to Caucasian-speaking groups (average  $F_{ST} = 0.019$ ) than to the other Altaic-speaking group (Turkey,  $F_{ST} = 0.099$ ). The  $D_A$ -values gave similar results. The population tree and PC plots (figure 4) also did not reflect the linguistic relationships; the Armenians and Azerbaijanians grouped consistently with their geographical neighbours in the Caucasus rather than with their respective Indo-European or Altaic linguistic neighbours.

In addition, we examined the correlation between the geographical distance and the genetic distance between pairs of Caucasus groups. The correlation between geographical distance and  $F_{ST}$  distance was not statistically significant when all Caucasus groups were included (Mantel test,  $\chi = 0.19$  and  $p = 0.165$  based on 1000 permutations). However, as discussed in § 3(d), the Ingushians appear to be an outlier amongst Caucasus populations, as also evidenced in the population tree (figure 4a) and  $F_{ST}$  (table 4) analyses. Removing the Ingushians resulted in a significant correlation between geographical distance and  $F_{ST}$  distance ( $\chi = 0.50$  and  $p < 0.01$ ), as shown in figure 5. Similar results were obtained when the  $F_{ST}$ -values were transformed as suggested by Rousset (1997) and for the correlation between geographical distance and the  $D_A$ -values (results not shown). Therefore, linguistic differences among the Caucasus groups appear not to have influenced genetic relationships as the significant correlation between geographical and genetic distance occurred across the different linguistic families. Moreover, the Caucasus Mountains have not been a substantial barrier to gene flow, as the significant correlation between geographical and genetic distance included comparisons between groups on either side of the mountains.

#### 4. DISCUSSION

The Caucasus groups were highly variable with respect to mtDNA HV1 sequences when compared with other European groups. The nucleotide diversity and mean

number of pairwise differences within groups were significantly higher on average for the Caucasus groups than for European groups (but lower than those which are observed in the Near Eastern groups). The mismatch distribution analysis and associated expansion times also indicated greater diversity within the Caucasus than in Europe and were consistent with the 'cline' in expansion times suggested by Calafell *et al.* (1996), starting with the oldest expansion times in Western Asia (65 000 years ago) and leading to the most recent expansion times in Europe (19 800 years ago). Moreover, the genetic differences between the Caucasus groups were larger on average than those between the European groups, as evidenced by the  $F_{ST}$ ,  $D_A$ , tree and PC analyses. This extensive mtDNA diversity in the Caucasus has not been revealed in previous mtDNA studies (Macaulay *et al.* 1999; Comas *et al.* 2000) due to the limited sampling of Caucasus groups. However, our mtDNA results are in excellent agreement with prior results based on analyses of classical genetic markers (Barbujani *et al.* 1994b) and *Alu* insertion polymorphisms (Nasidze *et al.* 2001).

The Caucasus groups also occupied an intermediate position between the European and Near Eastern groups in the population tree and PC plot. There are two possible explanations for this pattern: (i) Caucasus groups are derived from Near Eastern groups and are immediately ancestral to European groups, or (ii) Caucasus groups are admixed and have experienced gene flow from both Europe and West Asia. The first explanation thus corresponds to a strictly phylogenetic interpretation of the population tree and PC analyses, whereas the second explanation corresponds to interpreting these analyses strictly in terms of migration. Most analyses based on genetic distances among populations can be interpreted in either framework and, hence, are incapable of distinguishing between ancestry or gene flow (or a combination of the two). Phylogenetic analyses offer the possibility of distinguishing between these explanations, but mtDNA HV1 sequences are insufficient for accurate phylogenetic analysis; longer mtDNA sequences and additional loci will be required in order to resolve this issue.

A controversial hypothesis in linguistics is that the Caucasian and Basque languages are related, remnant pre-Indo-European languages (Gamkrelidze & Ivanov 1990; Ruhlen 1991) of palaeolithic antiquity. If so, one might expect to see evidence of a genetic relationship between Basque and Caucasus groups. However, the mtDNA results (figure 4) did not indicate any such relationship and the average  $F_{ST}$ - and  $D_A$ -values (table 4) were bigger between the Basque and Caucasus groups (0.026 and 0.075, respectively) than between the Basque and Indo-European groups (0.010 and 0.028, respectively). These results are in agreement with previous studies (Bertorelle *et al.* 1995; Comas *et al.* 2000).

At first glance, the high genetic diversity in the Caucasus might appear to be related to the high linguistic diversity. However, the genetic relationships among the Caucasus groups did not reflect linguistic relationships; grouping them based on various linguistic criteria did not change how the genetic variance was apportioned within versus between groups (table 3). Moreover, Caucasus populations tended to group with their geographical neighbours, not their linguistic neighbours (figure 4). This

pattern was noted previously in a comparison of Georgian and Kurdish mtDNA sequences (Comas *et al.* 2000), but our more extensive sampling of Caucasus groups provides further insights. In particular, we found that Armenians, who speak an Indo-European language, did not group with other Indo-European-speaking populations and Azerbaijanians, who speak an Altaic language, did not group with other Altaic-speaking populations. Rather, Armenians and Azerbaijanians grouped genetically with each other and then with other Caucasus groups.

It therefore appears as if the Armenian and Azerbaijani languages represent language replacements. Is there any discernible trace of a closer relationship between Armenians and Indo-Europeans than between other Caucasus groups and Indo-Europeans or between Azerbaijanians and Turkey than between other Caucasus groups and Turkey that would further support this hypothesis? The average  $F_{ST}$ - and  $D_A$ -values between Armenians and Indo-Europeans were 0.014 and 0.054, respectively, which indeed were slightly less than those for the other Caucasus groups and Indo-Europeans (average  $F_{ST}$  = 0.016 and  $D_A$  = 0.066), but the values between Azerbaijanians and Turkey were greater than those between Turkey and the other Caucasus groups ( $F_{ST}$  = 0.099 and 0.085 and  $D_A$  = 0.087 and 0.078, respectively).

Apparently, the language replacements in Armenia and Azerbaijan occurred with no detectable corresponding contribution of mtDNA types. One possible mechanism for such language replacements is the 'elite-dominance' process (Renfrew 1987). Under this model, a small number of Indo-European speakers may have moved into the territory of contemporary Armenia and Altaic-speakers into Azerbaijan and displaced the existing elite class, resulting in the concomitant replacement of the existing language. Historical evidence does support this model for both Azerbaijani and Armenian language origins. The Azerbaijani language was introduced via the spread of Altaic-speaking groups from the inner Eurasian steppes (Renfrew 1991). One such group of mounted pastoralists, the Oghuz, migrated to Azerbaijan around the 11th century (Johanson 1998). In the case of the Armenian language, a Near Eastern homeland near the historical territory of Armenia has been postulated for proto-Indo-European languages (Renfrew 1987; Gamkrelidze & Ivanov 1995). The spread of proto-Armenian into the Caucasus region is thought to be associated with the appearance of the Kura-Araxes culture in the southern Caucasus, 4500–5000 years ago (Gamkrelidze & Ivanov 1995).

The impact on the mtDNA composition of the existing population could have been negligible depending on the size of the incoming group, particularly if the incoming group was predominantly male. This hypothesis of language replacement in Armenia and Azerbaijan by male-mediated, elite dominance therefore predicts that the genetic relationships of these populations based on analyses of Y-chromosomal markers (which is in progress) should reflect their linguistic relationships more closely, depending on the extent of the genetic contribution of the incoming groups.

It may be that the lack of correspondence between linguistic and genetic relationships in the Caucasus is a

consequence of genetic drift due to isolation and/or a small population size. A large impact of genetic drift on Caucasus populations could explain the significant AMOVA results that were observed in this study, regardless of how the Caucasus populations are grouped linguistically (table 3). However, genetic drift would decrease genetic diversity within populations and the Caucasus mtDNA diversity was actually greater than the European mtDNA diversity. Moreover, extensive genetic drift should also erase any evidence of isolation by distance, whereas we did observe a significant correlation between geographical distances and genetic distances among the Caucasus groups (figure 5). Thus, the lack of any demonstrable effect of linguistic relationships on the genetic relationships of Caucasus populations is probably not due solely to genetic drift. The significant correlation between geographical and genetic distances is all the more remarkable in that it not only transcends language families, but it also holds across the Caucasus Mountains. However, it should be borne in mind that the sampling of the Caucasus groups in the present study, while the most extensive to date, may still have been too limited for detecting more subtle influences of geographical and/or linguistic barriers on the genetic structure of the Caucasus region.

We thank M. Kayser, H. Oota, S. Pääbo and B. Pakendorf for useful discussion and S. Clifford for technical assistance. This research was supported by a National Research Council COBASE grant to I.N. and M.S. and a grant from the National Science Foundation to M.S.

## REFERENCES

- Anderson, S. (and 13 others) 1981 Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.
- Aris-Brosou, S. & Excoffier, L. 1996 The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol. Biol. Evol.* **13**, 494–504.
- Barbujani, G., Nasidze, I. S. & Whitehead, G. N. 1994a Genetic diversity in the Caucasus. *Hum. Biol.* **66**, 639–668.
- Barbujani, G., Whitehead, G. N., Bertorelle, G. & Nasidze, I. S. 1994b Testing hypotheses on processes of genetic and linguistic change in the Caucasus. *Hum. Biol.* **66**, 843–864.
- Bertorelle, G., Bertranpetit, J., Calafell, F., Nasidze, I. S. & Barbujani, G. 1995 Do Basque-speaking and Caucasian-speaking populations share non-Indo-European ancestors. *Eur. J. Hum. Genet.* **3**, 256–263.
- Bertranpetit, J., Sala, J., Calafell, F., Underhill, P. A., Moral, P. & Comas, D. 1995 Human mitochondrial DNA variation and the origin of Basques. *A. Hum. Genet.* **59**, 63–81.
- Burckhardt, F., Von Haeseler, A. & Meyer, S. 1999 HvrBase: compilation of mtDNA control region sequences from primates. *Nucl. Acids Res.* **27**, 138–142.
- Calafell, F., Underhill, P., Tolun, A., Angelicheva, D. & Kalaydjieva, L. 1996 From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *A. Hum. Genet.* **60**, 35–49.
- Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A. & Bertranpetit, J. 1996 Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Mol. Biol. Evol.* **13**, 1067–1077.
- Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A., Bosch, E. & Bertranpetit, J. 1997 Mitochondrial DNA variation and the origin of the Europeans. *Hum. Genet.* **99**, 443–449.

- Comas, D., Calafell, F., Bendukidze, N., Fananas, L. & Bertranpetit, J. 2000 Georgian and Kurd mtDNA sequence analysis shows a lack of correlation between languages and female genetic lineages. *Am. J. Phys. Anthropol.* **112**, 5–16.
- Côrte-Real, H. B., Macaulay, V. A., Richards, M. B., Hariti, G., Issad, M. S., Cambon-Thomsen, A., Papiha, S., Bertranpetit, J. & Sykes, B. C. 1996 Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *A. Hum. Genet.* **60**, 331–350.
- Di Rienzo, A. & Wilson, A. C. 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl Acad. Sci. USA* **88**, 1597–1601.
- Felsenstein, J. 1993 *PHYLIP (phylogeny inference package)* v. 3.5c. Seattle, WA: Department of Genetics, University of Washington.
- Gamkrelidze, T. & Ivanov, V. 1990 The early history of Indo-European languages. *Sci. Am.* **262**, 110–116.
- Gamkrelidze, T. & Ivanov, V. 1995 *Indo-European and the Indo-Europeans*. Berlin: Mouton de Gruyter.
- Harpending, H. C., Sherry, S. T., Rogers, A. R. & Stoneking, M. 1993 The genetic structure of ancient human populations. *Curr. Anthropol.* **34**, 483–496.
- Johanson, L. 1998 The history of Turkic. In *The Turkic languages* (ed. L. Johanson & E. Csato), pp. 81–83. London: Routledge.
- Legendre, P., Lapointe, F.-J. & Casgrain, P. 1994 Modeling brain evolution from behavior: a permutational regression approach. *Evolution* **48**, 1487–1499.
- Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonne-Tamir, B., Sykes, B. & Torroni, A. 1999 The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am. J. Hum. Genet.* **64**, 232–249.
- Mountain, J. L., Hebert, J. M., Bhattacharya, S., Underhill, P. A., Ottolenghi, C., Gadgil, M. & Cavalli-Sforza, L. L. 1995 Demographic history of India and mtDNA-sequence diversity. *Am. J. Hum. Genet.* **56**, 979–992.
- Muskhelishvili, D. 1977 *The main problems of Georgian historical geography*. Tbilisi: Metsnierba Press.
- Nasidze, I., Risch, G. M., Robichaux, M., Sherry, S. T., Batzer, M. A. & Stoneking, M. 2001 *Alu* insertion polymorphisms and the genetic structure of human populations from the Caucasus. *Eur. J. Hum. Genet.* (In the press.)
- Nichols, J. 1997 The epicentre of the Indo-European linguistic spread. In *Archaeology and language. I. Theoretical and methodological orientations* (ed. R. Blench & M. Spriggs), pp. 122–148. London: Routledge.
- Piercy, R., Sullivan, K. M., Benson, N. & Gill, P. 1993 The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int. J. Legal Med.* **106**, 85–90.
- Redd, A. J., Takezaki, N., Sherry, S. T., McGarvey, S. T., Sofro, A. S. M. & Stoneking, M. 1995 Evolutionary history of the COII/tRNA-lys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol. Biol. Evol.* **12**, 604–615.
- Renfrew, C. 1987 *Archaeology and language*. London: Jonathan Cape.
- Renfrew, C. 1991 Before Babel: speculations on the origins of linguistic diversity. *Camb. Archaeol. J.* **1**, 13–23.
- Renfrew, C. 1992 Archaeology, genetics, and linguistic diversity. *Man* **27**, 445–478.
- Richards, M., Côrte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., Papiha, S., Hedges, R., Bandelt, H.-J. & Sykes, B. 1996 Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* **59**, 185–203.
- Rogers, A. R. & Harpending, H. 1992 Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–569.
- Rousset, F. 1997 Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* **145**, 1219–1228.
- Ruhlen, M. 1991 *A guide to the world's languages*. Stanford University Press.
- Schneider, S. & Excoffier, L. 1999 Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089.
- Schneider, S., Roessli, D. & Excoffier, L. 2000 *Arlequin v. 2.000: a software for population genetics data analysis*. Geneva, Switzerland: Genetics and Biometry Laboratory, University of Geneva.
- Tajima, F. 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.
- Vigilant, L., Pennington, R., Harpending, H., Kocher, T. D. & Wilson, A. C. 1989 Mitochondrial DNA sequences in single hairs from a southern African population. *Proc. Natl Acad. Sci. USA* **86**, 9350–9354.
- Ward, R. H., Frazier, B. L., Dew-Jager, K. & Pääbo, S. 1991 Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl Acad. Sci. USA* **88**, 8720–8724.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.